Hungry for Energy

"They won't be able to find enough electricity to run all the chips."

- Elon Musk

Insights

This report delves into the intricate relationship between AI, data centres, and electricity consumption. We will also look at some upcoming solutions and advancements in the infrastructure that supports this development in the world of chips.

The rapid evolution of Artificial Intelligence (AI) has propelled it to the forefront of numerous industries, including healthcare, finance, and technology. In Q1 2024, discussions regarding AI rose 17% quarter-over-quarter to 32% of earnings calls. This constitutes a new peak in mentions of AI in corporate earnings calls. Executives from nearly all verticals discussed how they are looking to integrate AI into their products and operations or how to prepare themselves for the age of AI. The CEO of British multinational oil and gas company Shell for example highlighted how AI assists their engineers in detecting anomalies remotely. The CEO of India-based IT services and consulting company Tech Mahindra announced that the company plans to upskill all IT staff in AI-related skills in the next financial year.

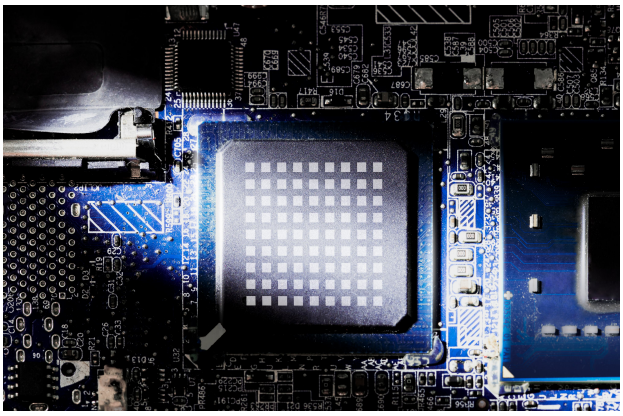
As AI systems become increasingly sophisticated (think Large Language models also known as LLMs, used in Generative AI systems), the demand for robust data processing and storage capabilities has surged. This demand has led to the proliferation of data centres, which are essential for handling the vast amounts of data generated and analysed by AI applications.

Data centres are notorious for their high energy consumption, a trend that is expected to continue as technology continues to pace ahead. The electricity consumed by data centres is primarily used to power servers and cooling systems. Maintaining optimal temperatures is crucial to prevent overheating and ensure the reliability and efficiency of data centre operations. Traditional cooling methods are energy-intensive, contributing significantly to the overall electricity consumption. This has risen critical concerns about sustainability and environmental impact.

In January this year, the International Energy Agency (IEA) issued its forecast for global energy use over the next two years. Included for the first time were projections for electricity consumption associated with data centres, cryptocurrency, and artificial intelligence. The IEA estimates that, added together, this usage represented almost 2 percent of global energy demand in 2022 — and that demand for these uses could double by 2026, which would make it roughly equal to the amount of electricity used by the entire country of 125 million in Japan. In the UK, the head of National Grid said in a speech in March, that data centre electricity demand in the UK will rise six-fold in just 10 years, fuelled largely by the rise of AI.

Chips market leader, Nvidia, highlighted the endless possibilities surrounding artificial intelligence. The company has unveiled plans for new AI accelerator chips, signalling its shift from graphics cards to AI chips. In March 2024, Nvidia unveiled the $70,000 Blackwell B200 GPU chip, the "world's most powerful AI chip." In less than three months, CEO Jensen Huang unveiled "Rubin", the successor to its "Blackwell". The unexpected move to reveal its next wave of products before Blackwell has even started shipping to customers shows how the world's most valuable chipmaker is racing to entrench its dominance of AI processors and the speed at which Chip technology is advancing.

The electricity market, confronted by substantial obstacles stemming from the gradual advancement of renewables and the pervasive electrification of daily essentials, anticipates noteworthy net demand increases due to emerging technologies. As semiconductor innovations progress rapidly, it is imperative to ensure that the accompanying infrastructure, pivotal for the success of these advancements, evolves in tandem, both in terms of expansion and capability. Let us take a look at what is happening in areas like electricity generation, enhancements in data centre efficiency, anticipated improvements in cooling systems, adoption of clean energy solutions for more sustainable electricity production, and forthcoming initiatives aimed at enhancing efficiency within this energy hungry sector.



1. The irony of more chips for more efficiency

Investing in energy-efficient hardware is a pivotal step towards superior performance per watt. Combining two processors on a personal computer is already standard practice for increasing computational power. This is now coming to data centres too. "We add a GPU, a $500 GPU, to a $1,000 PC, and the performance increases tremendously... when we do this in a data centre, a billion-dollar data centre, we add $500 million worth of GPUs, and all of a sudden, it becomes an AI factory" said Nvidia CEO Jensen Huang. Adding, "The more you buy, the more you save...". This is the CEO's math. Nvidia is ramping up to ship 15 million advanced AI server units per year over the next three years which are likely to consume energy equivalent to 225 million US households per year. These advancements however will enable data centres to handle AI workloads more efficiently, thereby reducing overall output per watt of consumption.

2. More efficient cooling

Innovative cooling solutions are essential for enhancing the energy efficiency of data centres. Cooling can account for up to 40% of a data centre's energy consumption.Techniques such as liquid cooling and use of phase change materials (PCMs) are significantly more efficient than traditional air-cooling methods. In liquid cooling, a coolant is run directly through microchannels in the chip, which in turn removes heat more effectively than the traditional use of fans. Microsoft implemented "sidekick" that sits next to the Maia 100 rack. These sidekicks work a bit like a radiator in a car. Cold liquid flows from the sidekick to cold plates that are attached to the surface of Maia 100 chips. Each plate has channels through which liquid is circulated to absorb and transport heat. That flows to the sidekick, which removes heat from the liquid and sends it back to the rack to absorb more heat, and so on.

In using PCMs, they absorb heat and change state (solid to liquid etc.) at certain temperatures. PCM provides passive thermal regulation, where no power is needed to create thermal regulation. There is no need for replacement as PCMs can be used repeatedly, promoting sustainability and eco-friendliness. There are many smaller players in this space. Honeywell's (HON - listed in America) thermal interface materials are based on proprietary technologies of polymer matrices and thermally conductive fillers, enabling them to handle challenging heat dissipation issues with long-term reliability and low cost of ownership.

3. Using AI for energy management

AI can be leveraged to optimize the energy efficiency of data centres. AI algorithms can predict and manage energy loads, ensuring that resources are utilized only when necessary and in the most efficient manner possible. Google's DeepMind, for instance, has implemented AI-driven cooling systems in its data centres, achieving a 40% reduction in energy used for cooling.

Digital Realty, a leading data centre platform of more than 300 data centres , is expanding its artificial intelligence platform for data centre energy monitoring, Apollo AI, into the Asia Pacific region. The company claims the AI platform has already demonstrated its effectiveness, identifying approximately 18GWh of expected and realized energy savings across the 16 sites where it is currently in use. Apollo uses machine learning to provide what Digital Realty describes as a "comprehensive dashboard" that lists optimization opportunities at each facility, prioritized by potential MWh savings.

> ***"The platform's self-learning capabilities aggregate findings, which are then audited and applied to newly onboarded facilities," the company said.***

4. Powering with clean energy

Transitioning to renewable energy sources is a critical strategy for reducing the carbon footprint of data centres. Leading tech companies like Google and Microsoft are at the forefront of this shift. Google, for instance, has committed to operating its data centres on carbon-free energy by 2030. By harnessing wind, solar, and hydroelectric power, these companies not only reduce their environmental impact but also mitigate energy cost volatility.

But, building as much renewable energy – wind and solar – in the near term has its limits. Each additional renewable has a higher incremental cost. At some point – say, 30% of penetration of renewables – we will likely reach a cliff. Nuclear energy – and Small modular nuclear reactors (SMR) in particular – can fill this gap that no renewable energy can.

Small modular reactors (SMRs) are fission-based systems that produce about one-third of the power generated by traditional nuclear plants — up to 300 megawatts of electricity, enough to power about 150,000 homes annually. They are also a fraction of the size, taking up about two soccer fields.

Microsoft Corp, partner of OpenAI which developed ChatGPT, has announced the future phases of its data centre development will be likely powered by co-located SMRs Its also recently announced the commencement of a US 100 billion data centre project (known as Stargate) to be deployed in the US by 2028. Global interest in SMRs is increasing due to their ability to meet the need for flexible power generation for a wider range of users and applications and replace ageing fossil fuel-fired power plants. They also display an enhanced safety performance through inherent and passive safety features, offer better upfront capital cost affordability and are suitable for cogeneration and non-electric applications. There are more than 80 SMR designs and concepts globally. Most of them are in various developmental stages and some are claimed as being near-term deployable. There are currently four SMRs in advanced stages of construction in Argentina, China and Russia.

5. Minimise latency with decentralize processing

Decentralizing data processing through edge computing can alleviate the burden on centralized data centres. By processing data closer to the source, edge computing reduces the energy required for data transmission and minimizes latency. This approach not only enhances efficiency but also improves the responsiveness of AI applications. This includes processing data on local devices or nearby computing resources, reducing dependence on distant data centres.

6. Policy and Regulation

Governments and regulatory bodies have a crucial role in promoting sustainable practices in data centre operations. Implementing standards for energy efficiency and providing incentives for the adoption of renewable energy can drive industry-wide changes. Policies that encourage transparency in energy usage and carbon emissions can also foster greater accountability and progress towards sustainability goals within the sector.

In conclusion, the rapid adoption of AI across various sectors is poised to have profound impacts, particularly on the energy sector, which are only beginning to be understood. Every online interaction relies on a vast network of information stored in remote servers, which require substantial energy to operate. According to the International Energy Agency (IEA), a single Google search consumes 0.3 watt-hours of electricity, whereas a request to ChatGPT uses 2.9 watt-hours. For context, an incandescent light bulb uses around 60 watt-hours. If ChatGPT were integrated into the 9 billion searches conducted each day, the IEA estimates that the resulting electricity demand would rise by 10 terawatt-hours annually—the equivalent consumption of about 1.5 million European Union residents.

While the enthusiasm for AI is understandable, it might be prudent to practice digital sobriety as responsible 21st-century consumers at least until energy infrastructure and efficiency catches up. This involves questioning the necessity of AI for mundane tasks, such as generating recipes or interacting with smart appliances. Sometimes, traditional methods are sufficient, and integrating AI might not always be necessary or beneficial.